

Predictive Complexity and Generalized Entropy Rate of Stationary Ergodic Processes

Mrinalkanti Ghosh and Satyadev Nandakumar

Department of Computer Science and Engineering,
Indian Institute of Technology Kanpur,
Kanpur, U.P., India.

Abstract. In the online prediction framework, we use generalized entropy to study the loss rate of predictors when outcomes are drawn according to stationary ergodic distributions over the binary alphabet. We show that the notion of generalized entropy of a regular game [10] is well-defined for stationary ergodic distributions. In proving this, we obtain new game-theoretic proofs of some classical information theoretic inequalities. Using Birkhoff's ergodic theorem and convergence properties of conditional distributions, we prove that a classical Shannon-McMillan-Breiman theorem holds for a restricted class of regular games, when no computational constraints are imposed on the prediction strategies.

If a game is mixable, then there is an optimal aggregating strategy which loses at most an additive constant when compared to any other lower semicomputable strategy. The loss incurred by this algorithm on an infinite sequence of outcomes is called its *predictive complexity*. We use our version of Shannon-McMillan-Breiman theorem to prove that when a restricted regular game has a predictive complexity, the predictive complexity converges to the generalized entropy of the game almost everywhere with respect to the stationary ergodic distribution.

1 Introduction

We consider the online prediction question studied by [15], [16], [10], [7], [9] in the setting of a stationary stochastic process. In this setting, we have a sequence of outcomes x_0, x_1, \dots from a finite alphabet. A predictor, given the history up to a certain index, predicts what the next outcome will be. We allow the predictor to present its prediction as a convex combination which represents the weight it assigns to each outcome in the alphabet. The game proceeds by revealing the next outcome, and then asking for the prediction of the future outcome. For an overview of this area, see [2]. Independently, Merhav and Feder [12], Feder [4] and Feder et. al. [5] have studied the question of optimal finite-state predictors with respect to Shannon entropy, in the setting of stationary Markov Chains. It is known that the log-loss game characterizes Shannon entropy. The present line of work generalizes their approach in two ways - first, in considering loss functions besides log-loss, and second, in considering optimal processes over stationary ergodic distributions.

A natural question in this context is how well the predictor is doing as the game progresses. We measure the discrepancy between the actual outcome and the predicted one, with a *loss function*. This helps us to ask whether *optimal* predictors exist - those which incur at most the same loss as any other predictor on any outcome, ignoring additive constants. Indeed if such an optimal predictor exists, we can use its loss rate on a particular sequence of outcomes to define its inherent *predictability* (see for example, [15], [16]).

Besides competitive advantage above other predictors, we can also characterize the performance of an optimal predictor by examining its expected loss assuming the outcomes are drawn from a particular distribution. Prior work by Kalnishkan et al. [10] establishes that if the outcomes are drawn independently according to a Bernoulli distribution on the alphabet, then the expected loss rate of an optimal predictor is the *generalized entropy* [8] of the loss function. In this paper, we extend this result to the important setting of stationary ergodic distributions.

The contributions of our paper are threefold.

1. First, we show that the generalized entropy rate of a stationary ergodic process is well-defined, if the game is *regular*. We provide “game-theoretic” proofs of classical information-theoretic inequalities, giving new intuitive proofs even in the special case of the Shannon entropy. This constitutes sections 3 and 4 of the paper.
2. Second, under a continuity and an integrability constraint, we show that optimal strategies exist for regular games.¹ We show that the loss rate incurred by such a strategy is the generalized entropy rate of the stationary ergodic process. This is a Shannon-McMillan-Breiman theorem for generalized entropy. This result is new, and we provide a proof using Vitali Convergence. This constitutes section 5 of the paper.
3. Using the above results, we show that when a game has *predictive complexity*, an optimal aggregator algorithm attains the entropy rate of the game. The proof that the aggregator incurs at most the entropy rate of loss crucially uses our Shannon-McMillan-Breiman Theorem. The proof that the aggregator incurs at least the entropy rate of loss uses some properties of stationary ergodic processes that we prove in Sections 3 and 4. This constitutes the final section of the paper.

2 Preliminaries

As defined in [10], a game \mathcal{G} is a triple $(\Sigma, \Gamma, \lambda)$ where Σ is a finite alphabet space, Γ is the space of predictions and $\lambda : \Sigma \times \Gamma \rightarrow [0, \infty]$ is the loss function, to be defined below. We will only consider the binary alphabet in this paper.

¹ There is an independent characterization of games with optimal strategies in terms of convexity of loss-regions [9]. We deal with this approach in the final section of our paper.

Intuitively, we model a predictor function which, given the string of outcomes so far, will predict the next outcome. We consider a slightly general framework where the predictor does not have to necessarily predict only one outcome. It is allowed to output a point $(p_0, p_1) \in I^2$ (equivalently, a probability vector, where p_0 is the predicted probability that the next bit is 0, and p_1 , the probability that the next bit is 1). The game proceeds by revealing the next outcome. Let this outcome be b . The prediction strategy is said to incur the loss $\lambda(b, (p_0, p_1))$.

As is customary, we adopt the notation \mathbb{N} for the set of natural numbers, starting from 0. The set of strings of length n is denoted Σ^n . The set of finite binary strings is denoted Σ^* and the set of infinite binary sequences is denoted Σ^∞ . For a finite or an infinite sequence x , the notation x_i^j denotes $x_i \dots x_j$. If x is shorter than n bits, x_0^{n-1} denotes x itself. If x is a finite string, and ω is a finite string or an infinite sequence, then $x \cdot \omega$ denotes the result of concatenating ω to x . For each natural number i , let Π^i be the class of all functions mapping i -long strings to I .

We call a family of functions \wp a *strategy* if $\forall i \in \mathbb{N}, |\wp \cap \Pi^i| = 1$, i.e, there is unique function which takes an i -length string as input and produce a strategy based on the input. We call that function \wp^i . Thus the prediction strategy is a non-uniform family. We impose no computational constraints until the final part of the paper.

3 Loss functions

The generalized entropy of a game is defined in terms of convex loss functions described above. We define the losses incurred by a strategy on a finite string w of outcomes, as the cumulative loss that it incurs on each bit of w . This follows the definition given in [10] and [9]. We generalize the notion slightly to deal with the expected loss that a strategy incurs with respect to a stationary distribution.

Definition 1. *The loss that a prediction strategy \wp , incurs on a finite string w of outcomes is defined to be*

$$Loss(w, \wp) = \sum_{i=0}^{|w|-1} \lambda(w_i, \wp^i(\omega_0^{i-1}))$$

In order to study when a strategy is better than another, we study the average loss it incurs, when outcomes are drawn from a stationary distribution. We consider the strategy which incurs the minimal expected loss on a particular set, if such a strategy exists. Let $(\Sigma^\infty, \mathcal{F}, P)$ be the probability space where \mathcal{F} is the Borel σ -algebra generated by cylinders

$$C_x = \{\omega \in \Omega \mid x \text{ is a prefix of } \omega\}$$

for all finite strings x . and $P : \mathcal{F} \rightarrow [0, 1]$ is the probability measure.

Let $X = (X_0, X_1, \dots)$ be a sequence of random variables on the probability space - for each $i \in \mathbb{N}$, X_i maps Σ^∞ to \mathbb{R} . For $k \geq 1$, let $S_k X$ denote the sequence (X_k, X_{k+1}, \dots) - that is, X "shifted left" k times.

Definition 2. [13] A sequence of random variables X is stationary if the probabilities of $S_k X$ and X coincide for every $k \geq 1$. That is, for every Borel set B in the σ -algebra over \mathbb{R}^∞ ,

$$P(X \in B) = P(S_k X \in B).$$

We could also use the terminology of measure-preserving transformations to capture stationarity. A transformation $T : \Omega \rightarrow \Omega$ is said to be *measure-preserving* if for every $A \in \mathcal{F}$, $P(T^{-1}A) = P(A)$. A measure-preserving transformation is said to be *ergodic* if $T^{-1}(A) = A$ if and only if $P(A)$ is either 0 or 1. [1]

The class of stationary processes correspond almost exactly to the class of probability spaces $(\Omega, \mathcal{F}, P, T)$ where $T : \Omega \rightarrow \Omega$ is a P -measure-preserving transformation. For $k \in \mathbb{N}$, let T^k denote the iterated application of T on itself, k times. It is easy to see that if T is measure preserving and X_0 is a random variable, then $(X_0, X_0 \circ T, X_0 \circ T^2, \dots)$ is a stationary sequence. We also have the converse.

Lemma 1. [13] For every stationary sequence X on a probability space (Ω, \mathcal{F}, P) , there is a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$, a random variable \tilde{X} and a \tilde{P} -measure preserving transformation $\tilde{T} : \tilde{\Omega} \rightarrow \tilde{\Omega}$ such that the distribution of $(\tilde{X}_0, \tilde{X}_0 \circ \tilde{T}, \tilde{X}_0 \circ \tilde{T}^2, \dots)$ coincides with the distribution of X .

On an alphabet space, we are interested in the coordinate random variables $X_i(\omega) = \omega_i$ ($i \in \mathbb{N}$), and any probability distribution such that X is stationary with respect to it, will be called a stationary distribution. A probability space with respect to which the left-shift transformation is ergodic will be called an *ergodic distribution*.

Definition 3. We define the n -step generalized entropy of the game to be

$$H_n = \inf_{\wp} \sum_{w \in \Sigma^n} P(w) \text{Loss}(w, \wp), \quad (1)$$

where $(\Sigma^\infty, \mathcal{F}, P)$ is a stationary probability space.

In order to avoid degenerate games (for example, games where the least expected loss is infinity, precluding any incentive to play the game), Kalnishkan et al.[10] restricts the game in the following manner.

- We restrict Γ to be a compact space. For the binary alphabet space, the prediction space is $[0, 1]$.
- The loss function λ is an extended real-valued convex function on $\Sigma \times \Gamma$. We take the discrete topology on the alphabet and the standard topology on $[0, 1]$. Then λ is continuous with respect to their product topology.
- There is a prediction $\gamma \in \Gamma$ such that for every $b \in \Sigma$, the inequality $\lambda(b, \gamma) < \infty$ holds. This property ensures that the n -ary entropy is a finite quantity.

- If there are $\gamma_0 \in \Gamma$ such that for some $b \in \Sigma$, the loss $\lambda(b, \gamma) = \infty$, then there is a sequence $\gamma_1, \gamma_2, \dots \rightarrow \gamma$ such that for each γ_i , we have $\lambda(b, \gamma_i) < \infty$.

A game which obeys these conditions is said to be *regular*. The last condition is necessary (but not sufficient) to ensure that predictive complexity exists for the game. We need this property crucially in Theorems 4 and 6.

The n step generalized entropy is the least expected loss incurred by any strategy, on Σ^n . Since Σ^n is a compact space and λ is continuous in both its arguments, the infimum in the above expression is attained by some strategy.²

Example 1. The Log-Loss game: Consider the binary alphabet and predictions be values in $[0,1]$. Let p_0 and p_1 be the probability of the bit 0 and bit 1, respectively.

Suppose we define the loss function by $\lambda(b, \gamma) = -\log(|\bar{b} - \gamma|)$, where b is a bit, \bar{b} its complement, and $\gamma \in [0,1]$. Then the minimal expected loss over one bit is obtained at $\gamma = p_1$, ensuring that $H(p_1)$ is the Shannon entropy of the distribution. (End of Example)

Definition 4. The generalized conditional entropy of Σ^n given Σ^m is defined as

$$\begin{aligned} H_{n|m} &= \inf_{\wp} \sum_{w \in \Sigma^m} P(w) \sum_{x \in \Sigma^n} P\{x | w\} \sum_{i=0}^{m-1} \lambda(x_i, \wp^{i+m}(w \cdot x_0^{i-1})) \\ &= \inf_{\wp} \sum_{wx \in \Sigma^{n+m}} P(wx) \sum_{i=0}^{m-1} \lambda(x_i, \wp^{i+m}(w \cdot x_0^{i-1})) \end{aligned}$$

This is an analogue of the definition of conditional Shannon entropy. The inner term in Definition 4 can also be expressed as follows.

$$\sum_{i=0}^{m-1} \lambda(x_i, \wp^{i+m}(w \cdot x_0^{i-1})) = \text{Loss}(wx, \wp) - \text{Loss}(w, \wp).$$

When we generalize the theory to handle arbitrary loss functions, we do lose some ideal properties that Shannon entropy has. The following theorem states that Shannon entropy is the unique function having certain ideal properties that we desire in a measure of information [11].

Theorem 1. Suppose F is a continuous function mapping n -dimensional probability distributions to $[0,1]$ having the following properties.

1. For any random variables A and B , $F(AB) = F(A) + F(B|A)$.
2. The n -dimensional uniform distribution has the largest entropy among n -dimensional distributions.
3. $F(p_1, p_2, \dots, p_n, 0) = F(p_1, p_2, \dots, p_n)$.

² The authors remark in [10] that such a strategy need not exist for Σ^* .

Then there is a positive constant λ such that for every n -dimensional probability vector (p_1, \dots, p_n) , $H(p_1, p_2, \dots, p_n) = \lambda F(p_1, p_2, \dots, p_n)$.

With our definition of the cumulative loss, we can establish the chain rule for generalized entropy.

Lemma 2. For all positive natural numbers m and n , we have $H_{m+n} = H_m + H_{n|m}$.

Proof. In Definition 4, \wp^i for $0 \leq i \leq m$ does not play any role in the infimum and likewise in Definition 3, \wp^i for $i \geq n$ does not play any role in the infimum. This observation allows us to deduce that

$$\begin{aligned} H_m + H_{n|m} &= \inf_{\wp} \left(\sum_{w \in \Sigma^m} P(w) \sum_{x \in \Sigma^n} P\{x | w\} \sum_{i=0}^{m-1} \lambda(x_i, \wp^{i+m}(w \cdot x_0^{i-1})) \right) + \\ &\quad \inf_{\wp} \sum_{w \in \Sigma^m} P(w) \text{Loss}(w, \wp) \\ &= \inf_{\wp} \sum_{w \in \Sigma^m} P(w) \left(\sum_{x \in \Sigma^n} \sum_{i=0}^{m-1} \lambda(x_i, \wp^{i+m}(w \cdot x_0^{i-1})) + \sum_{w \in \Sigma^m} \text{Loss}(w, \wp) \right). \quad (2) \end{aligned}$$

Now,

$$\begin{aligned} &\inf_{\wp} \sum_{w \in \Sigma^m} P(w) \left(\text{Loss}(w, \wp) + \sum_{w' \in \Sigma^n} P\{w' | w\} \sum_{i=0}^{m-1} \lambda(x_i, \wp^{i+m}(w \cdot x_0^{i-1})) \right) \\ &= \inf_{\wp} \sum_{w \in \Sigma^m} P(w) \sum_{w' \in \Sigma^n} P\{w' | w\} \left(\text{Loss}(w, \wp) + \sum_{i=0}^{m-1} \lambda(x_i, \wp^{i+m}(w \cdot x_0^{i-1})) \right) \\ &= \inf_{\wp} \sum_{w \in \Sigma^{m+n}} P(w) \text{Loss}(w, \wp) = H_{m+n}. \end{aligned}$$

Since λ is non-negative, it is clear that all entropies defined so far are non-negative. An immediate consequence of this is $H_{m+n} \geq H_m$ for all $m, n \geq 0$. We see that this style of proof referring to strategies in games yields new intuitive proofs of such inequalities.

Since conditions 1 and 3 in Theorem 1 are satisfied, Khinchin's uniqueness theorem therefore leads us to conclude that with a generalized entropy, the uniform distribution need not have maximal entropy - for example, the square-loss is not maximized at the uniform distribution.

4 Entropy of a Regular Game

The goal of this section is to define the notion of the entropy of a regular game. Our idea is to define it to be the limiting rate of the n -step generalized entropies of the game. We now show that if the game is regular and the probability distribution is stationary, such a limit exists. Thus the notion of the entropy of a regular game is well-defined.

Lemma 3. [Generalized Shannon Inequality] For any regular game and non-negative integers m and n , we have $H_{m/n} \leq H_m$.

Proof. The following proof is for $m = 1$. In this special case $H_1 = \inf_{\gamma \in \Gamma} \sum_{a \in \Sigma} P(a) \lambda(a, \gamma)$

and

$$H_{1/n} = \inf_{f \in \Pi^n} \sum_{w \in \Sigma^n} P(w) \sum_{a \in \Sigma} P\{a \mid w\} \lambda(a, f(w)) = \inf_{f \in \Pi^n} \sum_{a \in \Sigma} P(a) \sum_{w \in \Sigma^n} P\{w \mid a\} \lambda(a, f(w))$$

Now pick the $\gamma \in \Gamma$ which matches H_1 . We can do this because regularity condition of game requires Γ to be compact. The loss function is continuous in both its arguments ensuring that the expected loss in (1) is a continuous function on a compact space. Now define $f' : \Sigma^n \rightarrow \{\gamma\}$. Clearly, $f' \in \Pi^n$. So,

$$\begin{aligned} H_{1/n} &\leq \sum_{a \in \Sigma} P(a) \sum_{w \in \Sigma^n} P\{w \mid a\} \lambda(a, f'(w)) = \sum_{a \in \Sigma} P(a) \sum_{w \in \Sigma^n} P\{w \mid a\} \lambda(a, \gamma) \\ &= \sum_{a \in \Sigma} P(a) \lambda(a, \gamma) = H_1 \end{aligned}$$

The general case proceeds by induction by defining $f'^{i+n}(w w_0^{i-1}) = f^i(w_0^{i-1})$, where w is an n -long string and $1 \leq i \leq m$.

In the special case of the log-loss game with a Bernoulli distribution on the finite alphabet, the argument above yields a new argument for the Shannon inequality.

Lemma 4. For any regular game, any stationary distribution P defined on it, and any positive pair of natural numbers m and n , $H_{m/n} \geq H_{m/n+1}$.

Proof. We prove the inequality for $m = 1$. The general case would follow from application of Lemma 2. We have,

$$H_{1/n} = \inf_{f \in \Pi^n} \sum_{w \in \Sigma^n} P(w) \sum_{a \in \Sigma} P\{a \mid w\} \lambda(a, f(w)) = \inf_{f \in \Pi^n} \sum_{a \in \Sigma} \sum_{w \in \Sigma^n} P\{wa\} \lambda(a, f(w))$$

$$\text{and similarly } H_{1/n+1} = \inf_{f' \in F^{n+1}} \sum_{a \in \Sigma} \sum_{w \in \Sigma^{n+1}} P\{wa\} \lambda(a, f'(w)).$$

We show for each $f \in \Pi^n$ we have a $f' \in F^{n+1}$ which matches the inner quantity on which infimum is taken. Then, by taking infimum over F^{n+1} , we would have $H_{1/n} \geq H_{1/n+1}$. Fix a $f \in \Pi^n$ and consider $f' \in F^{n+1}$ defined as $f'(bw) = f(w)$ for all $w \in \Sigma^n, b \in \Sigma$. Now,

$$\begin{aligned} \sum_{a \in \Sigma} \sum_{w \in \Sigma^{n+1}} P\{wa\} \lambda(a, f'(w)) &= \sum_{a \in \Sigma} \sum_{b \in \Sigma} \sum_{w' \in \Sigma^n} P\{bw'a\} \lambda(a, f'(bw')) \\ &= \sum_{a \in \Sigma} \sum_{w' \in \Sigma^n} \sum_{b \in \Sigma} P\{bw'a\} \lambda(a, f(w')) \\ &= \sum_{a \in \Sigma} \sum_{w' \in \Sigma^n} P\{w'a\} \lambda(a, f(w')) \end{aligned}$$

where the last step follows from stationarity of P (i.e., $\sum_{b \in \Sigma} P\{bw\} = P\{w\}$ for all $w \in \Sigma^n$).

Theorem 2. *For any regular game \mathcal{G} and stationary $(\Sigma^\infty, \mathcal{F}, P)$, $\lim_{n \rightarrow \infty} \frac{H_n}{n}$ exists and is finite.*

Proof. From the regularity condition, we get H_1 is finite. From Lemma 2, it follows that $H_n = \sum_{i=0}^{n-1} H_{1|i}$.

By Lemma 4, $H_{1|k} \geq H_{1|(k+1)}$. Since entropies are non-negative, the sequence $\{H_{1|i}\}$ is a bounded, monotone decreasing sequence of reals. Hence, it has a limit which we denote by $H_{1|\infty}$. It also follows that $H_{1|\infty}$ is at most H_1 .

So by Cesàro mean, $\lim_{n \rightarrow \infty} \frac{H_n}{n} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} H_{1|i} = \lim_{n \rightarrow \infty} H_{1|n} = H_{1|\infty}$.

Definition 5. *Let $\mathcal{G} = (\Sigma^\infty, \Gamma, \lambda)$ be a regular game and $(\Sigma^\infty, \mathcal{F}, P)$ be a stationary distribution. Then The generalized entropy of the game is defined as*

$$H = \lim_{n \rightarrow \infty} \frac{H_n}{n}.$$

5 A Shannon-McMillan-Breiman Theorem

We now show that for regular games with a suitable restriction on the loss functions, optimal processes exist and they attain the generalized entropy rate of the stationary ergodic process. Our approach to this result is through uniform integrability and the Vitali Convergence theorem, which contrasts with the usual approach using the Dominated Convergence Theorem. First, we define the notion of a *strongly regular game*, for which the result holds.³ We will derive two consequences of strong regularity, *viz.*

1. The existence of a limiting function for the loss function, P -almost everywhere.
2. The integrability of this limiting function

We utilize these in the proof of the Shannon-McMillan-Breiman Theorem. We conclude with two examples, illustrating that Theorem 4 properly generalizes the classical Shannon-McMillan-Breiman theorem.

Definition 6. *Let (Ω, \mathcal{F}, P) be a probability space. A sequence of functions $\{f_n\}_{n=1}^\infty$ is called uniformly integrable if*

$$\lim_{\alpha \rightarrow \infty} \sup_n \int |f_n| I_{[|f_n| > \alpha]} dP = 0, \quad (3)$$

where $I_{[|f_n| > \alpha]}$ is the indicator function which is 1 at points ω with $|f_n(\omega)| > \alpha$ and is 0 otherwise.

³ Kalnishkan et al. [9] consider the notion of mixable games, which characterize regular games with optimality. In comparison, our conditions are based on integrability of the loss function.

If the sequence $\{f_n\}_{n=1}^\infty$ is uniformly integrable, then for every $\epsilon > 0$, and any large enough α ,

$$\sup_n \int |f_n| dP \leq \alpha + \epsilon \quad (4)$$

In addition to uniform integrability, we also need a continuity requirement over the space of strategies. We now introduce this. The next lemma characterizes $H_{1|n}$ in terms of the loss incurred by an optimal strategy on Σ^n .

Lemma 5.

$$H_{1|n} = \inf_{f \in \Pi^n} \sum_{w \in \Sigma^n} P(w) \sum_{a \in \Sigma} P\{a \mid w\} \lambda(a, f(w)) = \sum_{w \in \Sigma^n} P(w) \inf_{f \in \Pi^n} \sum_{a \in \Sigma} P\{a \mid w\} \lambda(a, f(w))$$

Proof. Let n be an arbitrary number. For any string w of length n , $P(w) \geq 0$, thus it follows that

$$\inf_{f \in \Pi^n} \sum_{w \in \Sigma^n} P(w) \sum_{a \in \Sigma} P\{a \mid w\} \lambda(a, f(w)) \geq \sum_{w \in \Sigma^n} P(w) \inf_{f \in \Pi^n} \sum_{a \in \Sigma} P\{a \mid w\} \lambda(a, f(w)),$$

hence it suffices to prove that the opposite inequality holds.

For each n -long string w , let f_w be the function which attains the infimum

$$\inf_{f \in \Pi^n} \sum_{a \in \Sigma} P\{a \mid w\} \lambda(a, f(w)).$$

Thus, the required expectation of infima can be written in terms of these functions as

$$\sum_{w \in \Sigma^n} P(w) \inf_{f \in \Pi^n} \sum_{a \in \Sigma} P\{a \mid w\} \lambda(a, f(w)) = \sum_{w \in \Sigma^n} P(w) \sum_{a \in \Sigma} P\{a \mid w\} \lambda(a, f_w(w)).$$

We can now define a function $f : \Sigma^n \rightarrow \Sigma$ as

$$f(w) = f_w(w), \quad w \in \Sigma^n.$$

It is clear from the definition of the function that

$$\sum_{w \in \Sigma^n} P(w) \sum_{a \in \Sigma} P\{a \mid w\} \lambda(a, f(w)) = \sum_{w \in \Sigma^n} P(w) \sum_{a \in \Sigma} P\{a \mid w\} \lambda(a, f_w(w)),$$

which implies the desired inequality.

Lemma 5 lets us analyse loss incurred by some “optimal” strategy. From Lemma 5, we can see given $w \in \Sigma^n$, optimal loss depends on the conditional probability distribution $(P\{0 \mid w\}, P\{1 \mid w\})$. Let $s(P\{0 \mid w\})$ be the strategy that gives optimal loss in $H_{1|n}$.

Let us define the following functions on Σ^∞ .

$$\begin{aligned} g_k(\omega) &= \lambda(\omega_0, s(P\{0 \mid \omega_{-k}^{-1}\})) \\ g(\omega) &= \lambda(\omega_0, s(P\{0 \mid \omega_{-\infty}^{-1}\})). \end{aligned}$$

$$\text{So, } \text{Loss}(\omega_0^{n-1}, \wp_n) = \sum_{k=0}^{n-1} g_k(T^k \omega).$$

Definition 7. A regular game is strongly regular if

1. s is a continuous function of the conditional probability.
2. For each natural number N , define $G_N : \Omega \rightarrow [0, \infty]$ by

$$G_N(\omega) = \sup_{k \geq N} |g_k(\omega) - g(\omega)|.$$

We require that $\{G_N\}_{N=1}^\infty$ is a uniformly integrable sequence.

First, we explain a consequence of condition (1). For a stationary ergodic distribution P , $P\{0 \mid \omega_{-k}^{-1}\} \rightarrow P\{0 \mid \omega_{-\infty}^{-1}\}$ as $k \rightarrow \infty$, and since g_k is a continuous function of the conditional distribution by condition (1), we have that $g_k \rightarrow g$ as $k \rightarrow \infty$, P -almost everywhere.

We now elicit some consequences of our assumption of uniform integrability. For uniformly integrable sequences of functions, their limit function is integrable even in the absence of any dominating function. This is known as the *Vitali Convergence Theorem* [6].

Theorem 3. Let (Ω, \mathcal{F}, P) be a probability space. If $\{f_n\}_{n=1}^\infty$ is a sequence of uniformly integrable functions such that $f_n \rightarrow f$ P -almost everywhere, then f is integrable and

$$\lim_{n \rightarrow \infty} \int |f_n - f| dP = 0.$$

Vitali Convergence of $\{G_N\}_{N=1}^\infty$ will be required in the final part of the proof of Theorem 4. We first show that uniform integrability of $\{G_N\}_{N=1}^\infty$ yields the integrability of the optimal loss.

Lemma 6. For a strongly regular game and a stationary distribution P ,

$$\lim_{n \rightarrow \infty} \int g_n dP = \int \lim_{n \rightarrow \infty} g_n dP = \int g dP.$$

Proof. We know that for each $n \in \mathbb{N}$,

$$\int |g_n| dP = \int g_n dP = H_{1|n},$$

which exists for regular games and stationary distributions. Now, for every n ,

$$\int |g_n| dP = \int |g - g_n - g| dP \geq \int |g| dP - \int |g - g_n| dP.$$

Hence we have

$$H = \lim_{n \rightarrow \infty} \int |g_n| dP \geq \int |g| dP - \liminf_{n \rightarrow \infty} \int |g - g_n| dP. \quad (5)$$

By the uniform integrability of $\{G_N\}_{N=1}^\infty$, we have that

$$\lim_{n \rightarrow \infty} \int |g - g_n| dP = 0.$$

Thus, by (5), we have $H \geq \int |g| dP$.

Using uniform integrability and the notion of continuity, we can introduce the setting for our Shannon-McMillan-Breiman Theorem.

For the sake of convenience, in the following proof, we will consider two-way infinite sequences. However, the same theorem holds for one-way sequences as well (see Chapter 13 of [1]). We briefly mention the formal correspondence.

Let (X, \mathcal{B}, μ) be a measure space with T being a measure preserving transform, not necessarily invertible. We construct a measure preserving system $(\hat{X}, \hat{\mathcal{B}}, \hat{\mu}, \hat{T})$ as follows.

- Define $\hat{X} = \{(x_i)_{i \in \mathbb{N}} \mid x_i \in T^{-i}X, Tx_{i+1} = x_i \text{ for all } i \in \mathbb{N}\}$
- Let $\pi_j : \hat{X} \rightarrow T^{-j}X$ be the projection function which projects j^{th} co-ordinate of an element of \hat{X} , i.e, $\pi_j(x) = x_j$. Construct a σ algebra \mathcal{B}' generated by sets of the form $\pi_i^{-1}T^{-i}E$, for all $i \in \mathbb{N}$, and $E \in \mathcal{B}$.
- Let $\hat{\mu}(\pi_i^{-1}T^{-i}E) = \mu(E)$ for all $E \in \mathcal{B}$.
- Complete \mathcal{B}' with respect to $\hat{\mu}$ to get $\hat{\mathcal{B}}$.
- Define $\hat{T} : \hat{X} \rightarrow \hat{X}$ by $\hat{T}((x_i)_{i \in \mathbb{N}}) = ((Tx_i)_{i \in \mathbb{N}})$.

Clearly, \hat{T} is an invertible transform given by $\hat{T}^{-1}(x_1, x_2, x_3, \dots) = (x_2, x_3, x_4, \dots)$. Since T is measure preserving, \hat{T} is also measure preserving. $(\hat{X}, \hat{\mathcal{B}}, \hat{\mu}, \hat{T})$ is called *natural extension of (X, \mathcal{B}, μ, T)* . It is ergodic iff the original system is ergodic. For unilateral alphabet system, its natural extension has same entropy. For details, see Fact 4.3.2 of [3].

Theorem 4. *For a strongly regular game $(\Sigma, \Gamma, \lambda)$, and stationary ergodic distribution $(\Sigma^\infty, \mathcal{F}, P)$, let H be the generalized entropy of the game. Moreover, let \wp be a strategy such that for every n , \wp^n achieves H_n . Then for $\omega \in \Omega$, the following holds:*

$$\lim_{n \rightarrow \infty} \frac{\text{Loss}(\omega_0^{n-1}, \wp^n)}{n} = H \quad (6)$$

for P -almost every ω .

We cannot use the Birkhoff's ergodic theorem (see for example, [1]) directly to prove the above theorem, since the summands in the Birkhoff average on the left of (6) depend in general on n , and are not the same integrable function. We however can use the convergence in conditional distributions ensured by a stationary distribution, in conjunction with Birkhoff's ergodic theorem to establish our result.

Proof. Recall that $g_k \rightarrow g$ almost everywhere, and $\int g$ exists by Lemma 6. We know $\text{Loss}(\omega_0^{n-1}, \wp^n) = g_n(\omega)$.

Since T is measure preserving transformation, by change of variable,

$$\int_{\Omega} g_k(\omega) dP = \int_{\Omega} g_k(T^k \omega) dP = H_{1|k}.$$

Thus

$$\int g(w) dP = \lim_{n \rightarrow \infty} \int g_n(w) dP = \lim_{n \rightarrow \infty} H_{1|n} = H.$$

By the Ergodic theorem, we get

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} g(T^k w) = \int g(w) dP = H,$$

for P -almost every $\omega \in \Omega$.

Now,

$$\frac{1}{n} \sum_{k=0}^{n-1} g_k(T^k w) = \frac{1}{n} \sum_{k=0}^{n-1} g(T^k w) + \frac{1}{n} \sum_{k=0}^{n-1} (g_k(T^k w) - g(T^k w)).$$

where the first term tends to H as $n \rightarrow \infty$. If we show second term in the previous equation is tends to 0 a.e. as $n \rightarrow \infty$, we are done.

Define $G_N(w) = \sup_{k \geq N} |g_k(w) - g(w)|$. By the assumption of strong regularity, the sequence of functions $\{G_N\}_{N=1}^{\infty}$ is uniformly integrable. Also, since $g_n \rightarrow g$ P -a.e., we know that $G_N \rightarrow 0$ P -almost everywhere as $N \rightarrow \infty$. By the Vitali Convergence Theorem,

$$\lim_{N \rightarrow \infty} \int G_N dP = \int \lim_{N \rightarrow \infty} G_N dP = 0.$$

Now for each N ,

$$\begin{aligned} \limsup_{n \rightarrow \infty} \left| \frac{1}{n} \sum_{k=0}^{n-1} (g_k(T^k \omega) - g(T^k \omega)) \right| &\leq \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} |g_k(T^k \omega) - g(T^k \omega)| \\ &\leq \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} G_N(T^k \omega) = \int G_N(\omega) dP \end{aligned}$$

where the last equality follows from Birkhoff Ergodic Theorem. Note that this holds for all values of N and right side converges to 0 a.e. as $N \rightarrow \infty$. Since the left side is non-negative, it is 0 a.e. So, $\frac{1}{n} \sum_{k=0}^{n-1} (g_k(T^k \omega) - g(T^k \omega)) \rightarrow 0$ as $n \rightarrow \infty$. This concludes the proof.

Recall that the generalized entropy of the log-loss game is the Shannon entropy. We now show the square loss and the log-loss games are strongly regular, thus establishing that we have a proper generalization of the classical Shannon-McMillan-Breiman theorem.

Example 2. Log-loss Game. The loss function $\lambda : \{0, 1\} \times [0, 1] \rightarrow [0, \infty]$ is defined by

$$\lambda(b, \gamma) = -\log(|b - \gamma|).$$

The optimal strategy is given by $s_k = P\{0 \mid \omega_{-k}^{-1}\}$, which is a continuous function of the conditional probability.

We have that for any N ,

$$\int \sup_{k \geq N} |g_k(\omega) - g(\omega)| dP \leq \int \sup_{n \geq 1} |g_n(\omega) - g(\omega)| dP \leq \int \sup_{n \geq 1} |g_n(\omega)| + \int g dP.$$

Hence to show that the sequence $\sup_{k \geq N} |g_k(\omega) - g(\omega)|$ is uniformly integrable, it suffices to show that

$$\int \sup_{n \geq 1} |g_n(\omega)| dP$$

is integrable. It is easy to show that for a stationary distribution P and any $r \in \mathbb{R}$,

$$P\{\omega \mid \sup_k |g_k(\omega)| \geq r\} \leq 2e^{-r},$$

from which the integrability of $\sup_k |g_k|$ follows.

Thus $\sup_{k \geq N} |g_k - g|$, for $N = 1, 2, \dots$ forms a uniformly integrable sequence of functions, and Theorem 4 holds for the log-loss game.

Example 3. Square-loss game. The loss function in the square loss game $\lambda : \{0, 1\} \times [0, 1] \rightarrow [0, 1]$ defined by

$$\lambda(b, \gamma) = (b - \gamma)^2. \tag{7}$$

The optimal strategy in the square-loss game is to pick $\gamma = p\{1 \mid \omega_{-k}^{-1}\}$, which is continuous in the conditional probability.

This loss function is bounded, hence

$$\int \sup_{k \geq 1} |g_k(\omega) - g(\omega)| dP \leq \int 1 dP = 1,$$

ensuring that $G_N = \sup_{k \geq N} |g_k(\omega) - g(\omega)|$ is uniformly integrable. Thus Theorem 4 holds for the square-loss game.

6 Predictive Complexity of Stationary Ergodic Games

We now consider computable prediction strategies. We would like to define the inherent unpredictability of a string x as the performance of an optimal computable predictor on x . It is not clear that one such predictor exists for any game. The work of Vovk and Watkins[15] establishes a sufficient condition for predictive complexity to exist.

Definition 8. A pair of points $(s_0, s_1) \in (-\infty, \infty]^2$ is called a *superscore*⁴ if there is a prediction $\gamma \in \Gamma$ such that $\lambda(0, \gamma) \leq s_0$ and $\lambda(1, \gamma) \leq s_1$. We denote the set of superscores for a regular game \mathcal{G} by \mathcal{S} .

Definition 9. A prediction strategy $\wp : \Sigma^* \rightarrow (-\infty, \infty]$ is called a *superloss process* if the following conditions hold.

1. $\wp(\Lambda) = 0$
2. For every string x , the pair $(\wp(x0) - \wp(x), \wp(x1) - \wp(x))$ is a superscore with respect to the game.
3. \wp is upper semicomputable.

A superloss process K is *universal* if for any superloss process \wp there is a constant C such that for every string x ,

$$K(x) \leq \wp(x) + C.$$

It follows that the difference in loss between any two superloss processes is bounded by a constant. Hence we may pick a particular superloss process \mathcal{K} and call $\mathcal{K}(x)$ the *predictive complexity* of the string x with respect to the game \mathcal{G} .

When we consider regular games, it is not necessary that an optimal strategy exists on Σ^* which incurs at most an additive loss when compared to any other prediction process. However, Vovk [14] and Vovk and Watkins[15] introduced the concept of *mixability* to ensure that one such universal process exists.

Definition 10. Let $\beta \in (0, 1)$. Consider the homeomorphism $h_\beta : (-\infty, \infty]^2 \rightarrow [0, \infty)^2$ specified by $h_\beta(x, y) = (\beta^x, \beta^y)$. A regular game \mathcal{G} with set of superscores \mathcal{S} is called β -mixable if the set $h_\beta(\mathcal{S})$ is convex. A game \mathcal{G} is called *mixable* if it is β -mixable for some $\beta \in (0, 1)$.

Theorem 5. [15] If a game \mathcal{G} with set of superscores \mathcal{S} is mixable, then \mathcal{G} has a predictive complexity.

It is known that the logloss and the square loss games are mixable. The coincidence of logloss and Kolmogorov complexity enables us to view predictive complexity as a generalization of predictive complexity. Absolute loss game is known not to be mixable [17].

We mention a loss bound which holds for mixable games. This is used in the proof of the theorem which follows.

Lemma 7. [10] If \mathcal{K} is predictive complexity of a mixable game \mathcal{G} , then there is a positive constant c such that $|\mathcal{K}(xb) - \mathcal{K}(x)| \leq c \ln n$ for all $n = 1, 2, \dots$, strings x and bits b .

We can now show that for a strongly regular mixable game \mathcal{G} , the predictive complexity rate on an infinite sequence of outcomes attains the generalized entropy of the stationary ergodic distribution P , almost everywhere.

⁴ In [10], [9], the concept is called a superprediction.

Theorem 6. Let $\mathcal{G} = (\Omega, \Gamma, \lambda)$ be a strongly regular mixable game with predictive complexity \mathcal{K} . Let (Ω, \mathcal{F}, P) be the probability space over the outcomes where P is a stationary ergodic distribution with generalized entropy H . Then

$$\lim_{n \rightarrow \infty} \frac{\mathcal{K}(\omega_0^{n-1})}{n} = H,$$

for P -almost every $\omega \in \Omega$.

Proof. (A) Upper Bound: First we show that $\lim_{n \rightarrow \infty} \frac{\mathcal{K}(\omega_0^{n-1})}{n} < H + \epsilon$ for any $\epsilon > 0$. This is an application of our Shannon-McMillan-Breiman theorem, Theorem 4 for generalized entropy.

Let \wp_n be the strategy which achieves $H_{1|n}$. There is a computable strategy ζ so that for all $0 \leq i \leq n-1$,

$$\lambda(a, \zeta_i(w)) < \lambda(a, \wp_n^i(w)) + \frac{\epsilon}{2}$$

for all $a \in \Sigma$ and for all $w \in \Sigma^i$. This is possible since set of all such strategies constitute an open set. By the definition of predictive complexity, we have

$$\begin{aligned} \mathcal{K}(\omega_0^{n-1}) &\leq \text{Loss}(w_0^{n-1}, \zeta) + O(1) \\ &\leq \text{Loss}(w_0^{n-1}, \wp_n) + \frac{\epsilon n}{2} + O(1) \end{aligned}$$

By the Shannon-McMillan-Breiman Theorem, for large enough n ,

$$\text{Loss}(w_0^{n-1}, \wp_n) + \frac{\epsilon n}{2} + O(1) \leq H + \epsilon O(n) + \frac{\epsilon n}{2} + O(1).$$

Taking limits as $n \rightarrow \infty$, we have that

$$\lim_{n \rightarrow \infty} \frac{\mathcal{K}(\omega_0^{n-1})}{n} < H + \epsilon.$$

(B) We now establish the reverse inequality, $\lim_{n \rightarrow \infty} \frac{\mathcal{K}(\omega_0^{n-1})}{n} > H - \epsilon$ for $\epsilon > 0$. Since

$$(K(\omega_0^{n-1} \cdot 0) - K(\omega_0^{n-1}), K(\omega_0^{n-1} \cdot 1) - K(\omega_0^{n-1}))$$

is a superscore, we have $E(\eta_n | \omega_0^{n-1}) \geq H_{1|n}$ where $\eta_n = K(\omega_0^{n-1}) - K(\omega_0^{n-1})$.

Now we can apply the martingale strong law of large numbers, Theorem VII.5.4 of [13] and get

$$\begin{aligned} \frac{K(\omega_0^{n-1})}{n} &= \frac{1}{n} \sum_{i=0}^{n-1} \eta_i = \frac{1}{n} \sum_{i=0}^{n-1} E(\eta_i | \omega_0^{i-1}) + o(1) \\ &\geq \frac{1}{n} \sum_{i=0}^{n-1} H_{1|n} + o(1) = H + o(1), \end{aligned}$$

where the last equality is obtained by Theorem 2.

Acknowledgments

The authors would like to thank John Hitchcock and Vladimir V'yugin for helpful discussions.

References

1. P. Billingsley. *Ergodic Theory and Information*. John Wiley & Sons, 1965.
2. N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, 2006.
3. T. Downarowicz. *Entropy in Dynamical Systems*. New Mathematical Monographs. Cambridge University Press, 2011.
4. M. Feder. Gambling using a finite state machine. *IEEE Transactions on Information Theory*, 37:1459–1461, 1991.
5. M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Transactions on Information Theory*, 38:1258–1270, 1992.
6. Gerald B. Folland. *Real Analysis*. Wiley, 1999.
7. L. Fortnow and J. H. Lutz. Prediction and dimension. *Journal of Computer and System Sciences*, 70:570–589, 2005.
8. P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *Annals of Statistics*, 32(4):1367–1433, 2004.
9. Y. Kalnishkan, V. Vovk, and M. V. Vyugin. Generalized entropies and asymptotic complexities of languages. In *Learning Theory, 20th Annual Conference on Learning Theory*, pages 293–307, 2007.
10. Yuri Kalnishkan, Volodya Vovk, and Michael V. Vyugin. Loss functions, complexities, and the legendre transformation. *Theor. Comput. Sci.*, 313(2):195–207, 2004.
11. A. Ya. Khinchin. *Mathematical Foundations of Information Theory*. Dover Publications, 1957.
12. N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.
13. A. N. Shiryaev. *Probability*. Graduate Texts in Mathematics v.95. Springer, 2 edition, 1995.
14. V. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, pages 153–173, 1998.
15. V. G. Vovk and Chris Watkins. Universal portfolio selection. In *COLT*, pages 12–23, 1998.
16. Michael V. Vyugin and Vladimir V. V'yugin. Predictive complexity and information. In *COLT*, pages 90–104, 2002.
17. Vladimir V'yugin. Suboptimal measures of predictive complexity for absolute loss function. *Information and Computation*, 175:146–157, 2006.